The Dialog of Primary and Non-primary Auditory Cortex at the 'Cocktail Party'

CellPress

The Dialog of Primary and Non-primary Auditory Cortex at the 'Cocktail Party'

Elia Formisano[1,2,3,*] and Lars Hausfeld[1,2]
[1]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, PO Box 616, 6200 Maastricht, the Netherlands
[2]Maastricht Brain Imaging Centre, 6200 Maastricht, the Netherlands
[3]Maastricht Centre for Systems Biology, 6200, Maastricht, the Netherlands
*Correspondence: e.formisano@maastrichtuniversity.nl
https://doi.org/10.1016/j.neuron.2019.11.031

In this issue of *Neuron*, O'Sullivan et al. (2019) measured electro-cortical responses to "cocktail party" speech mixtures in neurosurgical patients and demonstrated that the selective enhancement of attended speech is achieved through the adaptive weighting of primary auditory cortex output by non-primary auditory cortex.

Imagine listening to the news while your son watches his favorite YouTube videos in the same room or listening to the exciting presentation of an interesting poster at a busy session, surrounded by many other presenters and attendants. In such cases, intentionally directing attention to a certain speaker biases the processing of the sound mixture at the ears, such that relevant speech is selectively enhanced and processed, and all other irrelevant sounds are suppressed. In this issue of *Neuron*, O'Sullivan et al. (2019) investigate the neural mechanisms enabling this selective enhancement of relevant speech in noisy scenes using electro-cortical recordings (ECoGs) in neurosurgical patients as they listened to multi-talker speech. Using a similar combination of multi-talker stimuli and ECoG, previous studies had shown that selectively attending to one speaker enhances the neural representation of that speaker in auditory cortex (AC). In these previous studies, however, recordings were limited to the superior temporal gyrus (STG), an anatomical region encompassing several non-primary auditory areas. Hence, the neural processing and attention effects in primary auditory areas, located along the Heschl's gyrus (HG), could not be investigated. In this new study, O'Sullivan et al. (2019) measured neural responses in both HG and STG combining depth (stereotactic electroencephalogram in HG) and surface (subdural ECoG) recording techniques. In this way, they examined the distinct contribution of primary and non-primary auditory cortical regions to the processing of multi-talker scenes as well as their hierarchical relation.

The main finding of this new study is that selective attention to a speaker modulates responses in STG and only to a limited extent in HG, where responses are instead selective for one or the other speaker irrespective of the locus of attention. This is clearly illustrated in the article's Figure 1C, where exemplary high-gamma responses to single-talker stimuli (i.e., the audiobooks spoken by speaker 1 [Spk1, male] or speaker 2 [Spk2, female] in isolation) are compared with the responses to multi-talker stimuli (i.e., the mixture of the audiobooks) for one electrode in STG and one in HG. In STG, the responses to the multi-talker stimuli changed depending on which speaker is being attended to and closely resembled the responses to the corresponding speaker presented in isolation. In contrast, the multi-talker responses in HG changed only weakly and resembled, for this electrode, the response to Spk1 presented in isolation irrespective of the attentional focus. Subsequently, the authors assessed these observations quantitatively by defining two indices, the "speaker selectivity index" (SSI), which measures the difference of response amplitude to single-talker stimuli, and the "attention modulation index" (AMI), which measures the effect of attention on the similarity between responses to single-talker and multi-talker stimuli. The SSI was found to be largest in electrodes located along HG (Figure 2), whereas the AMI was largest in STG sites (Figure 3); locations with high SSI and those with high AMI (Figure 4) were clearly dissociated.

Hence, findings in O'Sullivan et al. (2019) point to the modulation of responses in STG as the most evident and robust neural signature of selectively attending to a target speaker in a multi-talker speech mixture. Obtained with direct cortical recordings, these results provide compelling confirmation and validation of studies that measured brain responses to multi-talker speech using non-invasive electro- and magneto-encephalography (MEG) (e.g., Ding and Simon, 2012; Hausfeld et al., 2018). Especially, attention effects reported in MEG studies were large and most significant at around 150 ms but small and not significant at earlier latencies. This is consistent with the strong attention modulation of STG responses at 150 ms and, at least partially, with the weak (but this time significant) modulation of HG responses peaking around 80 ms, as reported in O'Sullivan et al. (2019). The analysis of the dependence of STG and HG multi-talker responses on masking levels in the mixtures further reinforces the agreement between invasive and non-invasive studies (Figure 5, O'Sullivan et al., 2019). In STG, the relation between multi-talker responses to the to-be-attended speaker and the corresponding single-talker responses is unaffected by masking levels of the to-be-ignored speaker. In contrast, in HG, masking levels of the to-be-ignored speaker do affect the relation between multi-talker and single-talker responses. Similarly, MEG responses at ~50 ms have been reported to be sensitive to the level of masking, whereas responses at ~120 ms to be invariant to masking for a broad range of signal-to-noise ratios (SNRs).

Converging evidence from invasive and non-invasive studies thus supports the interpretation that, when listening to mixtures of multiple speakers, HG responses mainly reflect the acoustic (energy-based) analysis of both the attended and the unattended speech streams, whereas STG responses reflect processing of the attended stream at a more abstract level. Understanding the representations and computational mechanisms subserving speech processing in STG is the focus of much current research (Yi et al., 2019). Model-based analysis of data as those collected by O'Sullivan et al. (2019) will help discern which acoustic-to-linguistic transformations most accurately account for the observed responses in STG.

The absence (or reduced strength) of top-down effects in primary auditory cortex (PAC) raises an interesting question on the role of PAC during active listening. Other studies investigating the effects of context, task, or attention on the neural analysis of sounds did observe significant modulations of activity in PAC/HG (e.g., Rutten et al., 2019). Neuronal populations in PAC/HG adapt their sensitivity to task-relevant acoustic features, enhancing the processing of relevant stimuli (King et al., 2018).

So, why are attention effects in PAC/HG small or absent during multi-talker speech? The specific processing demands that stimuli and tasks require may provide an answer to this question. In O'Sullivan et al. (2019), the two speech streams overlapped spectro-temporally but also markedly differed in their fundamental frequency (in Spk 1, F0 = 65 Hz; in Spk2, F0 = 175 Hz). Contrary to the reported effects, a selective enhancement of the attended pitch was to be expected, e.g., in lateral HG sites (Bendor and Wang, 2006), as it would contribute to segregating the speakers. Given the pitch difference and SNR at which the speech streams were mixed and presented, however, listeners may have had clean "views" of both speakers even at the highest masking levels. Furthermore, participants were asked to report the last (attended) sentence and thus they may have exploited higher-level linguistic (contextual, phonological, syntactic, and semantic) information, superseding acoustic information when this latter was noisy. That speakers could be accurately decoded from HG responses to the mixtures (Figure 6, O'Sullivan et al., 2019) and that listeners performed the task without much difficulty (average performance level = 90%) both lend support to this possibility. To further elucidate the role of HG during active listening, in future studies it will be important to examine the responses to a larger variety of speech mixtures (e.g., varying the range of masking as well as the pitch differences between speakers) and under different task requirements.

The recording of responses from both HG and STG enabled O'Sullivan et al. (2019) to investigate the relation between these regions. The characteristics of HG (faster, acoustically selective, weak attentional modulation) and STG (slower, less acoustically selective, strong attentional modulation) responses clearly pointed to a hierarchical model, with STG receiving input from HG. In line with this hypothesis, linear predictions of STG responses from HG responses were more accurate than those of HG responses from STG responses. Interrogation of the STG prediction model showed that the "weighting" of HG sites was dependent on the attended speaker and stronger for sites with higher speaker selectivity. Interestingly, an additional analysis showed that grouping of speaker-selective sites could be determined in an unsupervised manner based on the temporal coherence of neural responses. These results put forward the modulation of neural connectivity from HG to STG, possibly achieved through the adaptation of synaptic efficiency, as a potential neural mechanism enabling the selective enhancement of the attended speaker (and suppression of the unattended speaker). Depending on attentional demands, the throughput to STG from HG sites coherently responding to the attended speaker is increased while that of HG sites coherently responding to the unattended speaker is decreased.

These findings represent an important step toward the mechanistic description of sound analysis in AC. Several additional pieces, however, are still missing to reveal the intricate puzzle of neural information processing within the auditory cortical network. First, HG and STG are macro-anatomical regions comprising multiple (primary and non-primary) auditory areas, whose functional properties differ substantially. A finer anatomical differentiation of HG/STG locations, e.g., along the caudal-rostral axis (Jasmin et al., 2019), is needed to investigate the distinct contribution of different locations to the processing of specific components of speech. Second, neural responses in the (high) gamma frequency range, as examined in O'Sullivan et al. (2019), are known to reflect the feedforward processing of neural information. Additional analyses of responses in the lower frequency bands are needed to gain a more complete view of feedback neural signaling and modulatory processes involved (Scheeringa and Fries, 2019). Third, the flow of neural information for feedforward and feedback processing travels along spatially segregated channels across the cortical layers (Scheeringa and Fries, 2019). Thus, important advancements in understanding inter-areal communication are expected from measurement techniques that preserve the layer specificity of neural responses. Direct electrophysiological recordings with laminar electrodes can provide signals at the exquisite spatiotemporal resolution required to model neural processing at a layer-specific level and inter-laminar interactions. However, these recordings are invasive and with limited brain coverage; they can thus be complemented with non-invasive techniques, such as laminar fMRI, which provides neuro-vascular responses at a sub-millimeter resolution and, potentially, whole-brain coverage (De Martino et al., 2018). Whereas laminar fMRI does not allow examination of neural responses at a single-layer level, there is accumulating evidence that its spatial specificity is sufficient for distinguishing the major feedforward and feedback processing pathways. Ultimately, it will be the convergence of results from these invasive and non-invasive techniques that will unravel the full pattern of directed functional interactions within the network of AC areas and between AC and the rest of the brain.

REFERENCES

Bendor, D., and Wang, X. (2006). Cortical representations of pitch in monkeys and humans. Curr. Opin. Neurobiol. *16*, 391–399.

De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludağ, K., De Weerd, P., Ugurbil, K., Goebel, R., and Formisano, E. (2018). The impact of ultra-high field MRI on cognitive and computational neuroimaging. Neuroimage *168*, 366–382.

Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.

Hausfeld, L., Riecke, L., Valente, G., and Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. Neuroimage *181*, 617–626.

Jasmin, K., Lima, C.F., and Scott, S.K. (2019). Understanding rostral-caudal auditory cortex contributions to auditory perception. Nat. Rev. Neurosci. *20*, 425–434.

King, A.J., Teki, S., and Willmore, B.D.B. (2018). Recent advances in understanding the auditory cortex. F1000Res. *7*, 7.

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G.M., Sheth, S.A., Mehta, A.D., and Mesgarani, N. (2019). Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. Neuron *104*, this issue, 1195–1209.

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., and Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. Nat. Hum. Behav. *3*, 974–987.

Scheeringa, R., and Fries, P. (2019). Cortical layers, rhythms and BOLD signals. Neuroimage *197*, 689–698.

Yi, H.G., Leonard, M.K., and Chang, E.F. (2019). The Encoding of Speech Sounds in the Superior Temporal Gyrus. Neuron *102*, 1096–1110.